Forecast Calibration and Combination

Tom Hamill

tom.hamill@noaa.gov

www.cdc.noaa.gov/people/tom.hamill

Acknowledgments: Jeff Whitaker, CDC

Definitions

- **“Calibration”**
  - Synonymous with reliability; summarizes the conditional probability of observed | forecast.
  - More broadly, think of as post-processing to correct bias, spread deficiencies, sampling error, or to generate PDFs for non-state variables (e.g., stream flow, wave height, heating-degree days).
  - Desired result: maximal sharpness given perfect reliability. Happy customers.
- **“Combination”**
  - Synthesizing probabilities when provided with forecasts from multiple, independent forecast systems. Presumably similar desired result.

Disadvantages to calibration?

- Calibration research doesn't correct the underlying problem. Prefer to achieve unbiased, reliable forecasts by doing numerical modeling correctly in the first place.
  - Forecasts may be improved, but to end products not raw forecasts, so little gain in meteorological insight.
- Corrections may be model-specific; the calibrations for GFS v 2.0 may not be useful for ECMWF, much less GFS v 3.0.
- Could constrain model development. Calibration ideally based on long database of prior forecasts (reforecasts, or hindcasts) from same model. Do we delay model upgrades until new set of reforecasts completed?
- Complicated calibration methods may be difficult to maintain.
- Not that much is gained through calibration (at

Advantages to calibration?

(My assumption: calibration based on a large database of reforecasts from the same model.)

- Large gains in forecast skill may be possible, equivalent to 5-10 years of NWP development. [More later]
- Reforecast database required for calibration useful in model development. Can help detect subtle biases present only in large samples, e.g., biases in extreme weather forecasts. NWP

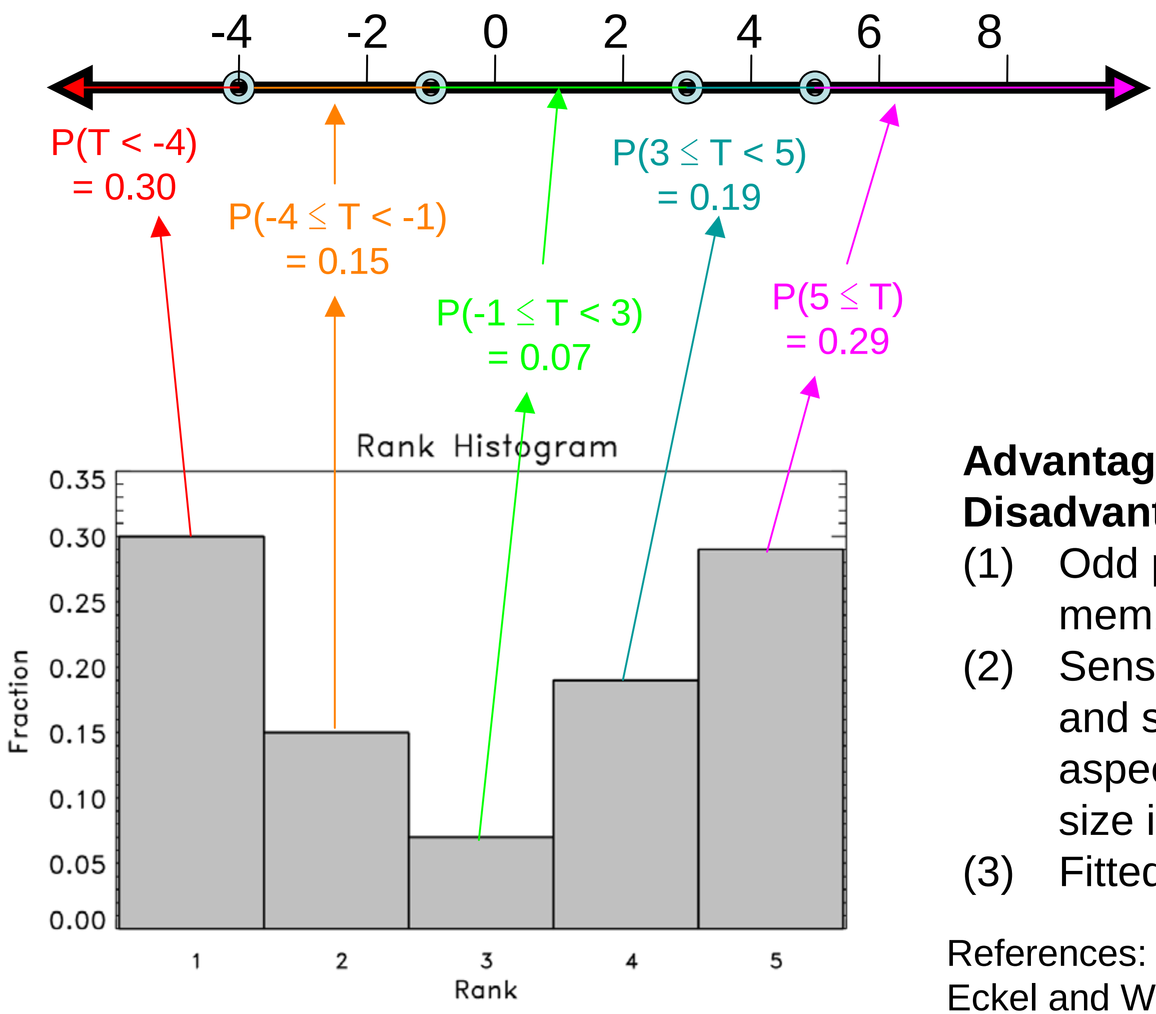
  developers are not used to utilizing reforecasts, so they don’t know what they’re missing.
- Calibration and model development can co-exist if NWP centers adopt dual track procedure, with reforecasts done every few years with lower-resolution version of model
- With dual-track, costs of reforecasts manageable
  - Reforecast computation can be distributed to other non-production computers.
  - Our work suggests that most of ensemble information contained in the mean; therefore, large-member reforecast ensembles surprisingly unnecessary.
- Maintenance issues not so bad if same model used unchanged, year after year.

A very brief review of calibration: (1) Model Output Statistics ("MOS")

```
KBID   GFS MOS GUIDANCE    2/16/2005  1800 UTC
 DT /FEB  17                  /FEB  18                /FEB  19
 HR   00 03 06 09 12 15 18 21 00 03 06 09 12 15 18 21 00 03 06 12 18
 N/X              32          40          25          35       19
 TMP  42 39 36 33 32 36 38 37 35 33 30 28 27 30 32 31 28 25 23 19 27
 DPT  34 29 26 22 19 18 17 17 17 17 17 15 14 13 11  8  7  6  5  2  4
 CLD  OV FW CL CL SC BK BK BK BK BK BK BK SC BK BK BK BK FW CL CL CL
 WDR  26 30 32 32 32 31 29 28 30 32 31 31 31 31 30 29 31 32 33 33 27
 WSP  12 12 12 11 08 08 09 08 09 09 10 10 10 12 13 13 15 16 15 09 08
 P06        17     0     0     0     4     0    10     6     8  0  0
 P12              17           0          10          17        8
 Q06         0     0     0     0     0     0     0     0     0  0  0
 Q12               0           0           0           0        0
 T06      0/ 2  0/ 0  1/ 0  1/ 2  0/ 1  0/ 1  1/ 0  0/ 1  0/ 0  0/ 0
 T12                  1/ 0        1/ 2        1/ 1        0/ 1  0/ 0
 POZ   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 POS  13 47 70 84 91100 96100100100100 92100 98100100100 94 92100100
 TYP   R  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S
 SNW                                       0                    0
 CIG   7  8  8  8  8  8  8  8  8  7  7  7  8  7  7  7  8  8  8  8  8
 VIS   7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7
 OBV   N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N
```

US: Statistical corrections to operational US NWS models, some fixed (NGM), some not (Eta, GFS). Refs: http://www.nws.noaa.gov/mdl/synop/index.htm, Carter et al., *WAF*, **4**, p 401, Glahn and Lowry, *JAM*, **11**, p 1580. **Canadian** models discussed in Wilson and Vallee, *WAF*, **17**, p. 206, and *WAF*, **18**, p 288. **Britain**: Met Office uses "updateable MOS" much like perfect prog.

Ensemble calibration: rank histogram techniques



NCEP MRF precipitation forecasts, from Eckel and Walters, 1998

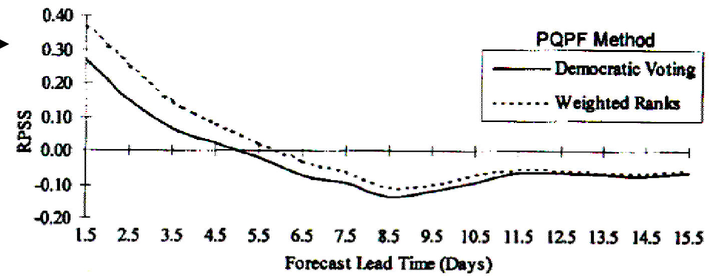


FIG. 10. Ranked probability skill score (RPSS) results for all forecast lead times.

Advantages: Demonstrated skill gain

Disadvantages:

(1) Odd pdfs, especially when two ensemble members close in value.

(2) Sensitive to shape of rank histogram, and shape of histogram may vary with aspects like precip amount --> sample size issues.

(3) Fitted parametric distributions as skillful

References: Hamill and Colucci (*MWR*, 1997, 1998; Eckel and Walters, *WAF*, 1998; used at UKMO)

Fitting parametric distributions

SMOOTHING OF FORECAST ENSEMBLES 2827

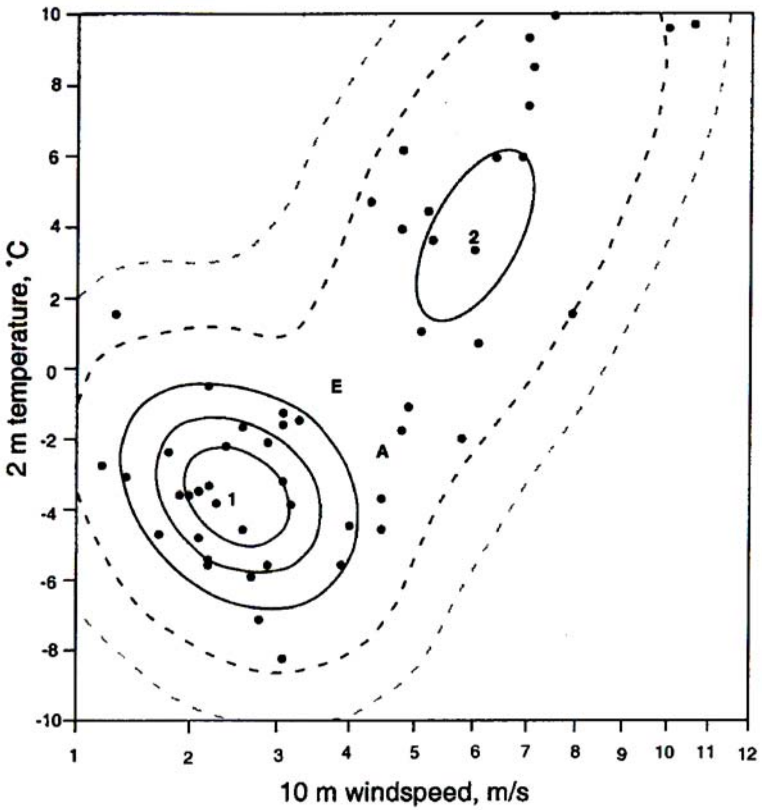
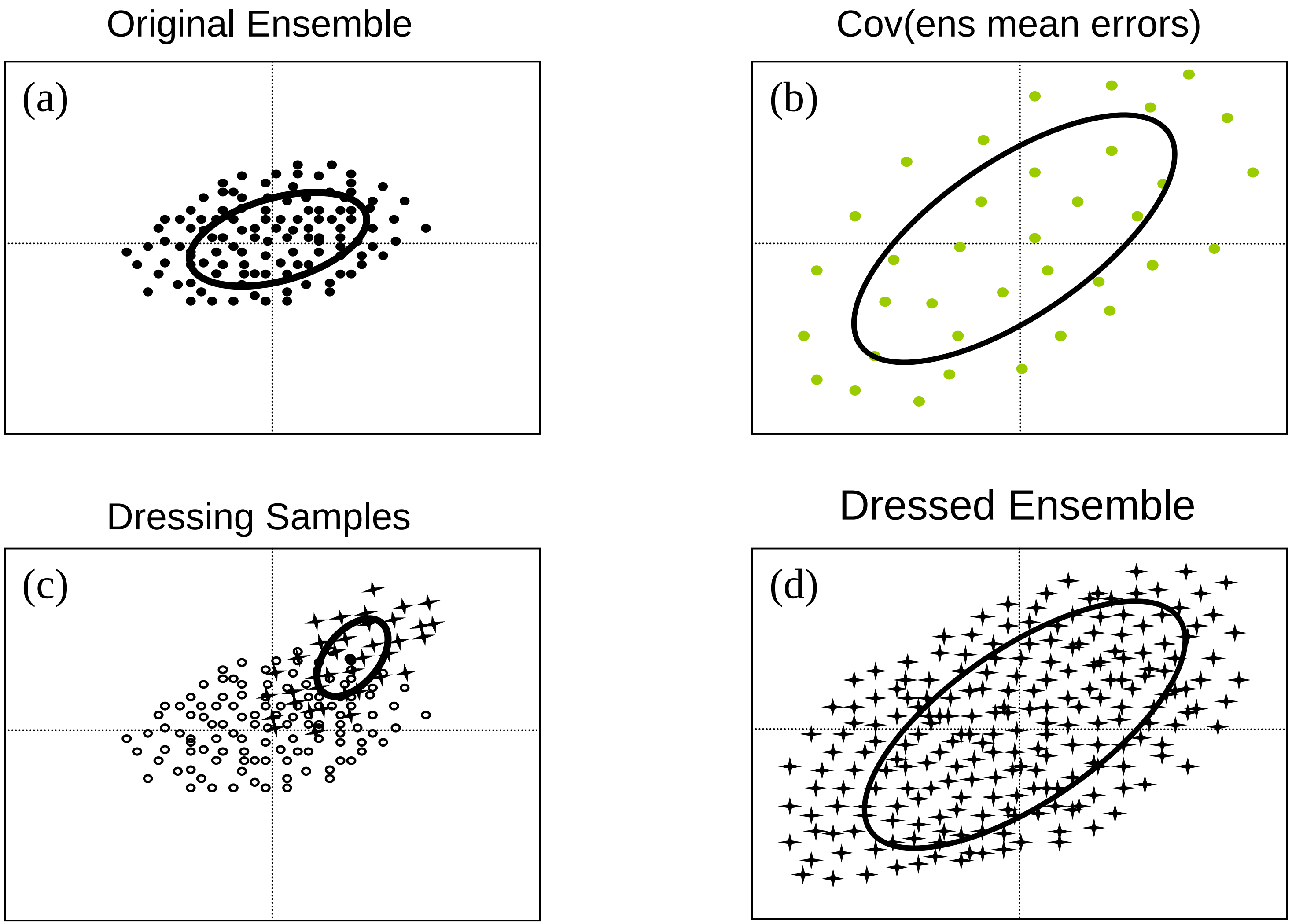


Figure 2. Example ensemble distribution with fitted Gaussian mixture, jointly for the temperature and wind-speed forecast at 12 UTC 10 January 1997 at Manchester, made at the 180 h lead time. Dots indicate individual forecasts made by the 51 ensemble members, with the ensemble mean located at 'E'. The two bivariate Gaussian densities $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are centred at '1' and '2', respectively, and the smooth lines indicate level curves of their mixture $f_{\mathrm{mix}}(\mathbf{x})$, formed with $\alpha = 0.57$ (see text). Contour interval is 0.05, and the thick and thin dashed lines are for 0.01 and 0.001, respectively. Subsequent verifying analysis is 'A'.

Wilks (*QJRMS*, **128**, p 2821) explored fitting parametric distributions, or mixtures thereof, to ECMWF forecasts in perfect-model context. Power-transformed non - Gaussian variables prior to fitting. Didn't address ensemble model errors in this study.

Dressing methods



Method of correcting spread problems. Assume prior bias correction.

Adv: Demonstrated improvement in ETKF ensemble forecasts in NCAR model.

Dis: Only works if too little spread, not too much.

Ref: Roulston and Smith (Tellus, 55A, p 16); Wang and Bishop (QJRMS, submitted; picture above)

Bayesian Model Averaging (BMA)

$$p(y \mid f_1, \ldots, f_K) = \sum_{k=1}^{K} w_k \; g_k(y \mid f_k)$$

Weighted sum of kernels centered around individual, bias-corrected forecasts.

Advantages: Theoretically appealing.

Disadvantages: [My personal opinion]: For Raftery's application (post-processing U. Washington MM5 ensemble), method over-fit training data. Shown here, with small sample, BMA radically de-weighted some members due to co-linearity. Expect this wouldn't happen when trained with larger sample.

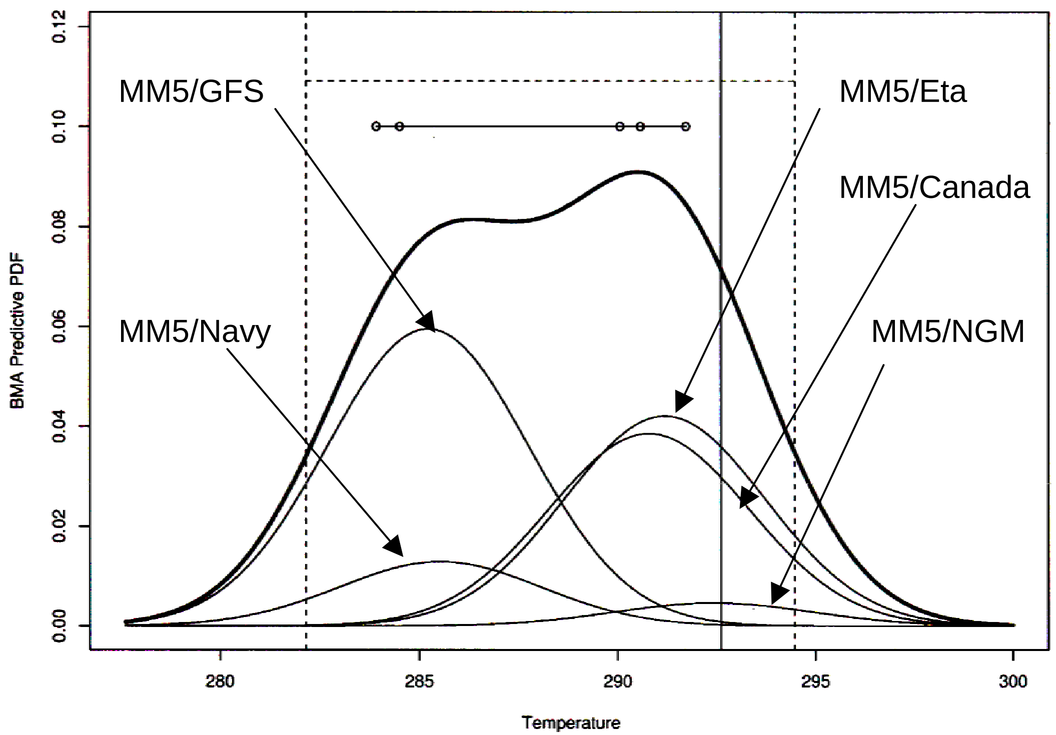


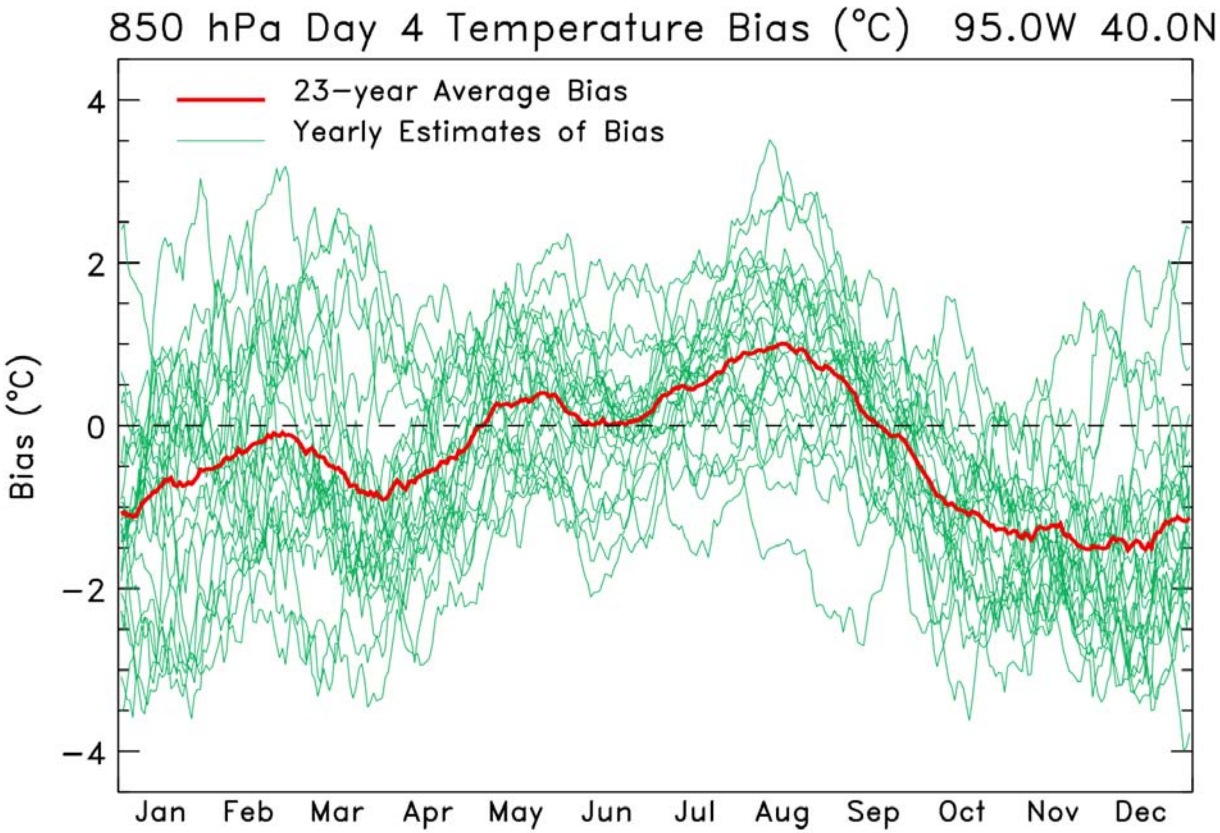
Figure 3: BMA predictive PDF (thick curve) and its five components (thin curves) for the 48-hour surface temperature forecast at Packwood, Wash., initialized at 0000 UTC on June 12, 2000. Also shown are the ensemble member forecasts and range (solid horizontal line and bullets), the BMA 90% prediction interval (dotted lines), and the verifying observation (solid vertical line).

Ref: Raftery et al., MWR, in press. Also recent work by Wilson at Canadian MSC.

A tool for exploring calibration: the CDC MRF reforecast data set

- **Definition of "reforecast"** : a data set of retrospective numerical forecasts using the same model to generate real-time forecasts.
- **Model**: T62L28 NCEP MRF (now "GFS"), circa 1998 (*http://www.cdc.noaa.gov/people/jeffrey.s.whitaker/refcst* for details).
- **Initial states**: NCEP-NCAR reanalysis plus 7 +/- bred modes (Toth and Kalnay 1993).
- **Duration**: 15-day runs every day at 00Z from 19781101 to now. (*http://www.cdc.noaa.gov/people/jeffrey.s.whitaker/refcst/week2).*
- **Data**: Selected fields (winds, geo ht, temp on 5 press levels, and precip, t2m, u10m, v10m, pwat, prmsl, rh700, conv. heating). NCEP/NCAR reanalysis verifying fields included (Web form to download at *http://www.cdc.noaa.gov/reforecast*).
- **Experimental PQPF**: *http://www.cdc.noaa.gov/reforecast/narr/*

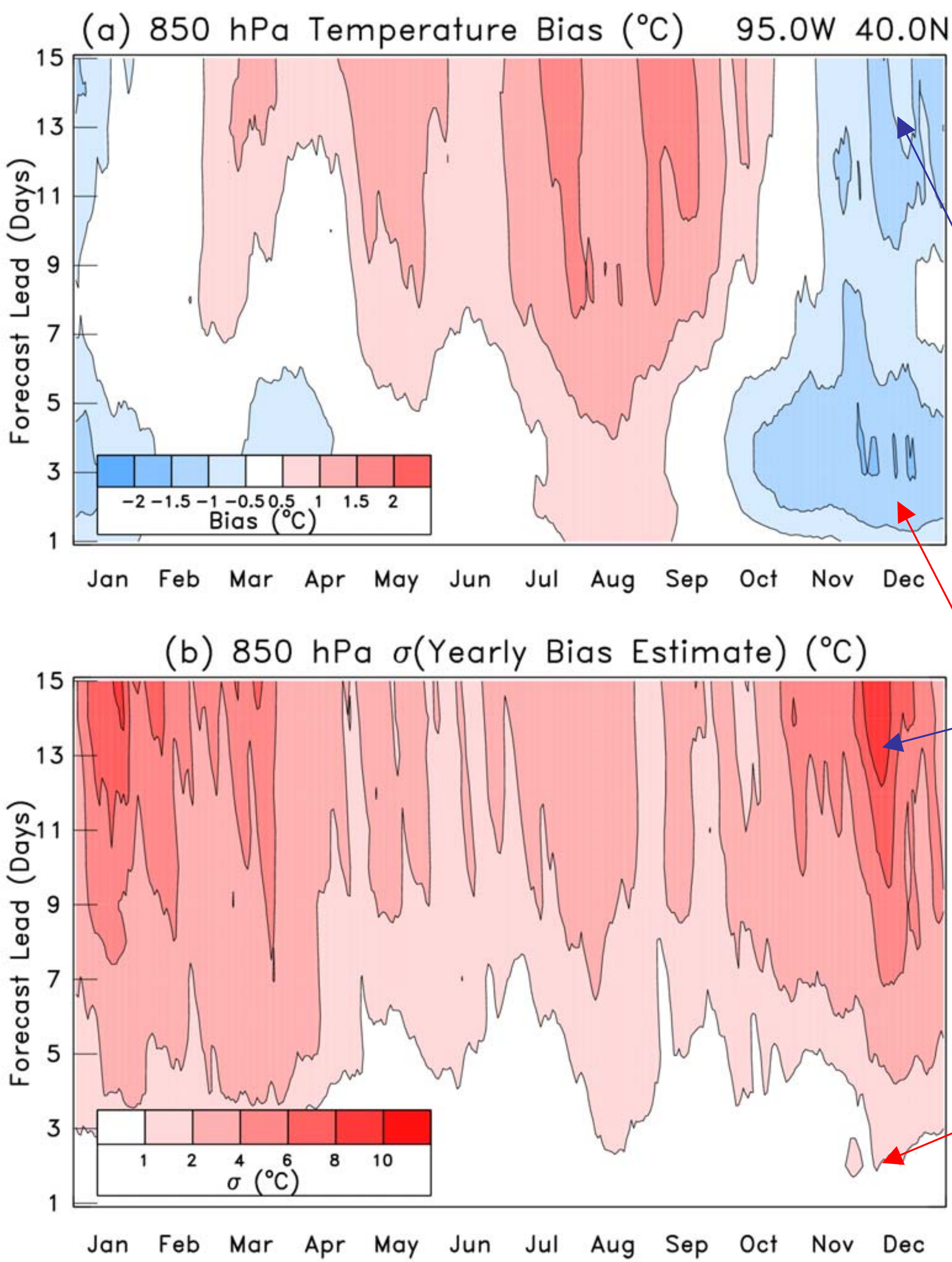
Issues arising in calibration. (1) Are large reforecast data sets *really* necessary?



Red curve shows bias averaged over 23 years of data (bias = mean F-O in running 61-day window)

Green curves show 23 individual yearly running-mean bias estimates

Note large inter-annual variability of bias.



When are long reforecast data sets necessary, and when are they not?

Example: bias correction.

Here, **large training data set required**; bias is small relative to its yearly variability.

Here, **small training data set adequate**; bias comparable or greater than its yearly variability.

Sample size in analog forecast technique

Step 2: find dates of old analogs

Step 3: extract observed weather

Forecast analog 1, 2/12/95

Observed Wx, 2/12/95

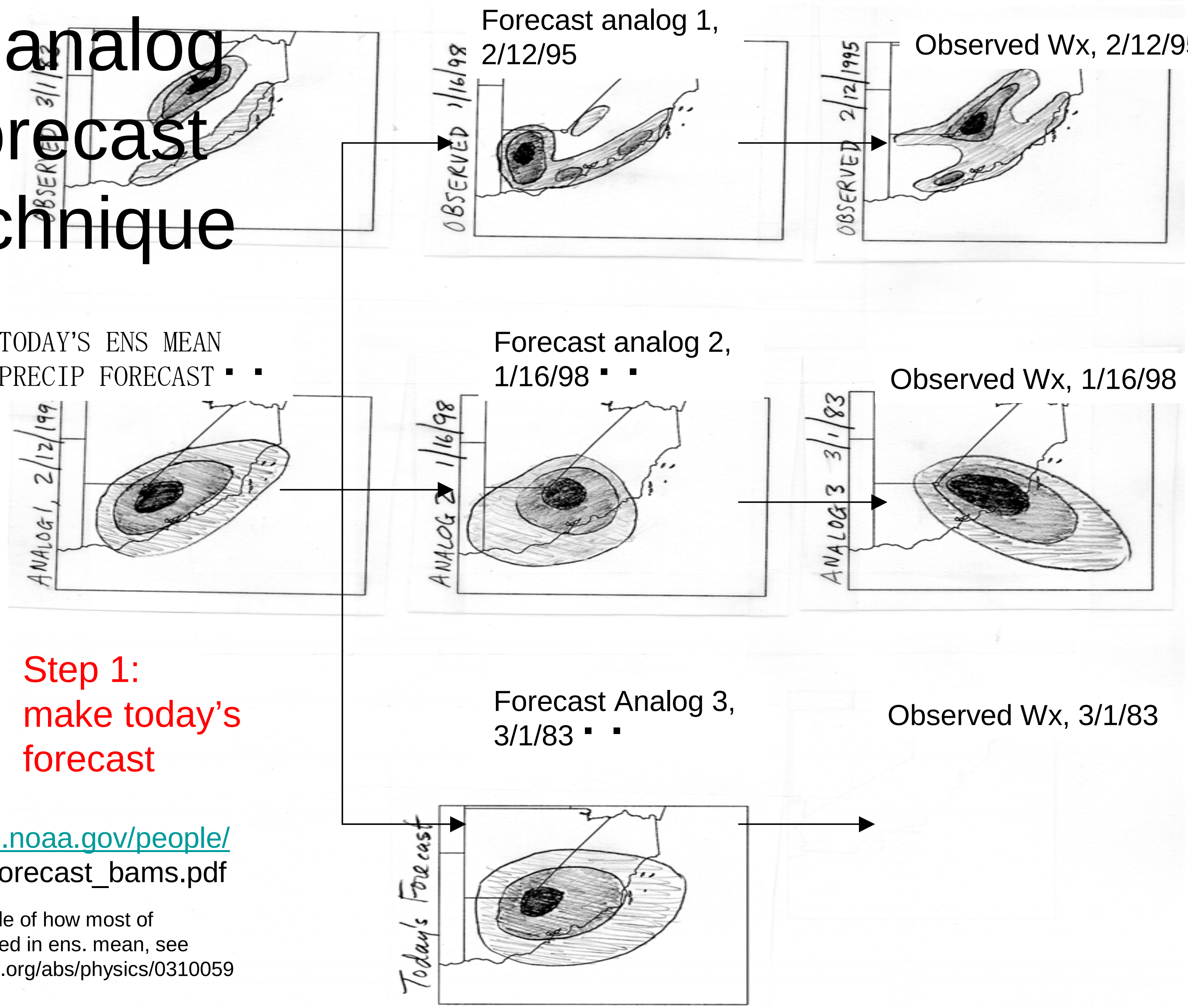
TODAY'S ENS MEAN PRECIP FORECAST . .

Forecast analog 2, 1/16/98 . .

Observed Wx, 1/16/98

Forecast Analog 3, 3/1/83 . .

Observed Wx, 3/1/83

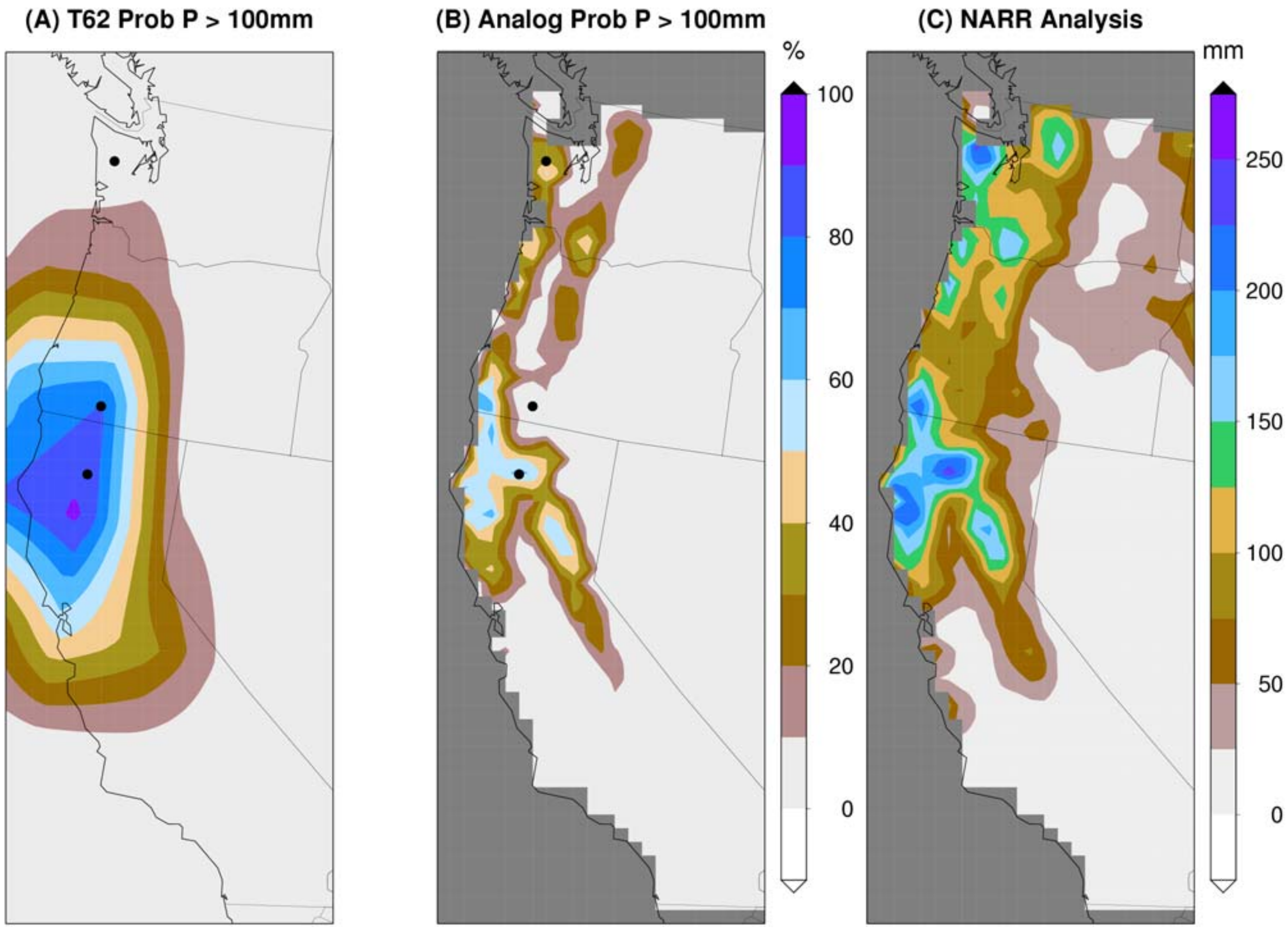


Step 1: make today's forecast

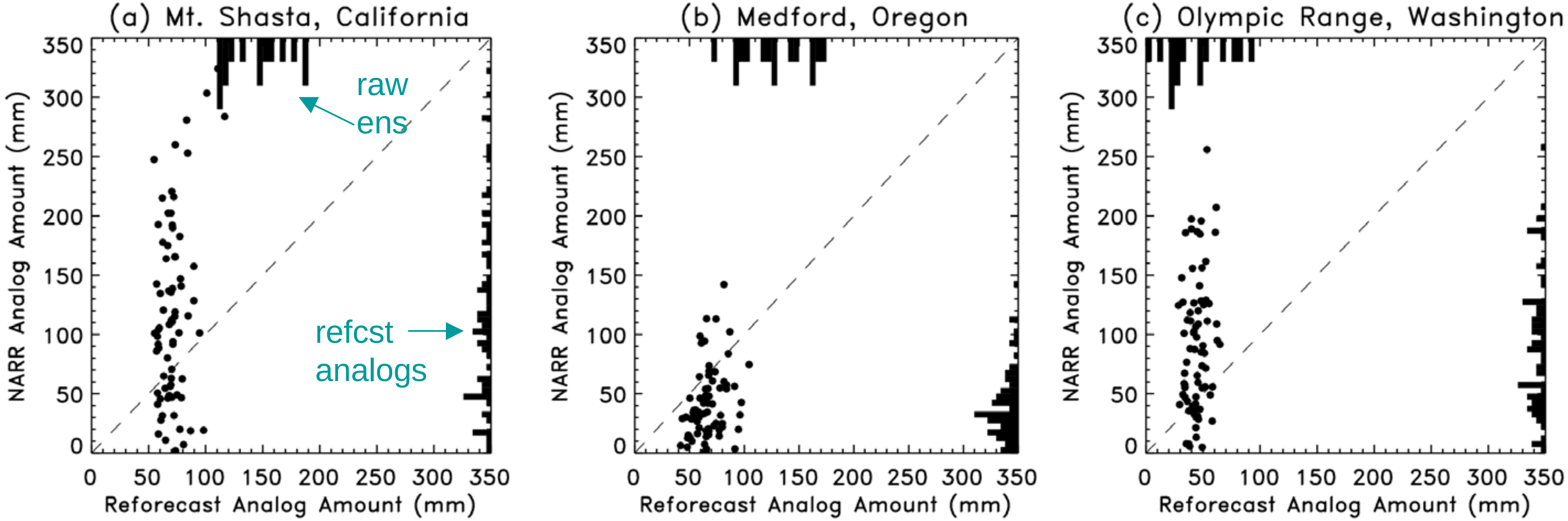
Ref: www.cdc.noaa.gov/people/ tom.hamill/reforecast_bams.pdf

For another example of how most of information contained in ens. mean, see Jewson, http://arxiv.org/abs/physics/0310059

Analog example:
Day 4-6 heavy precipitation in California, 0000 UTC 29 December 1996 - 0000 UTC 1 January 1997



Downscaling through analogs



Can't find any other reforecast analogs with precip as heavy. But introduce large scatter by taking associated observed analogs.

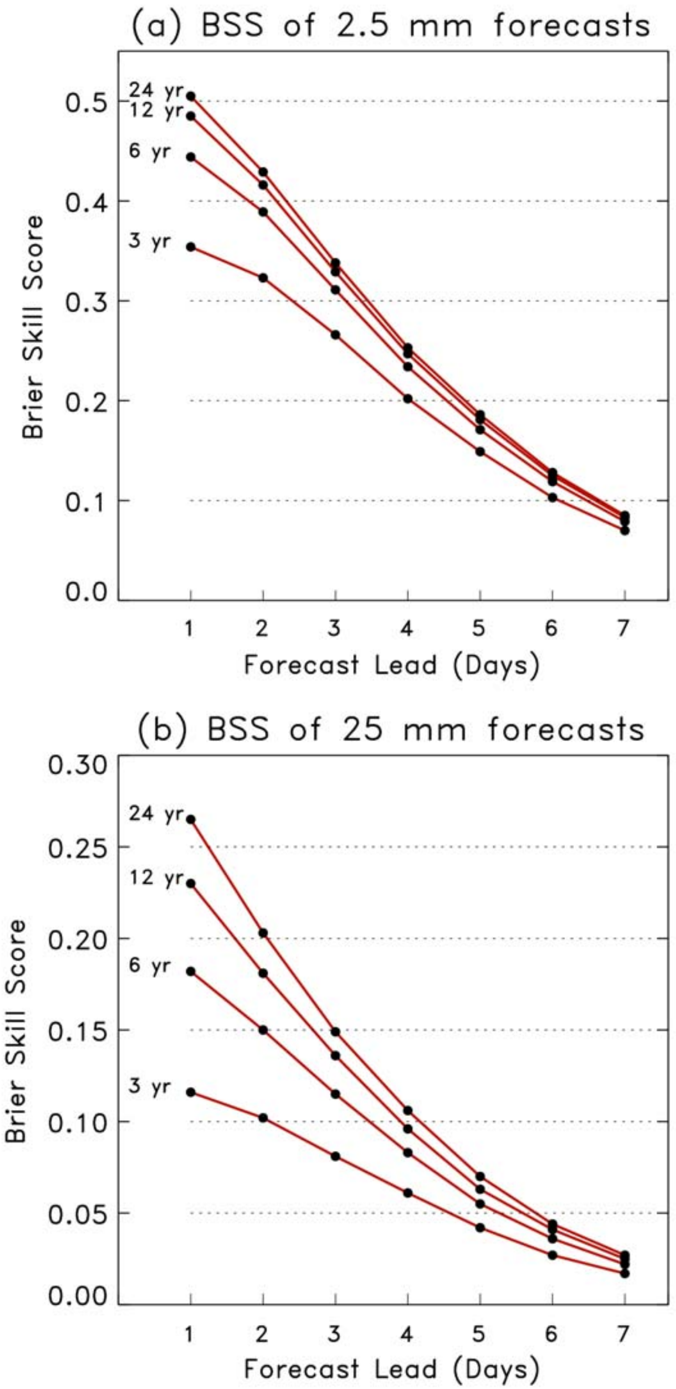
Again, few close reforecast analogs. But observed data recognizes overforecast bias.

Here there are close reforecast analogs. Observed data introduces spread, increases amount.

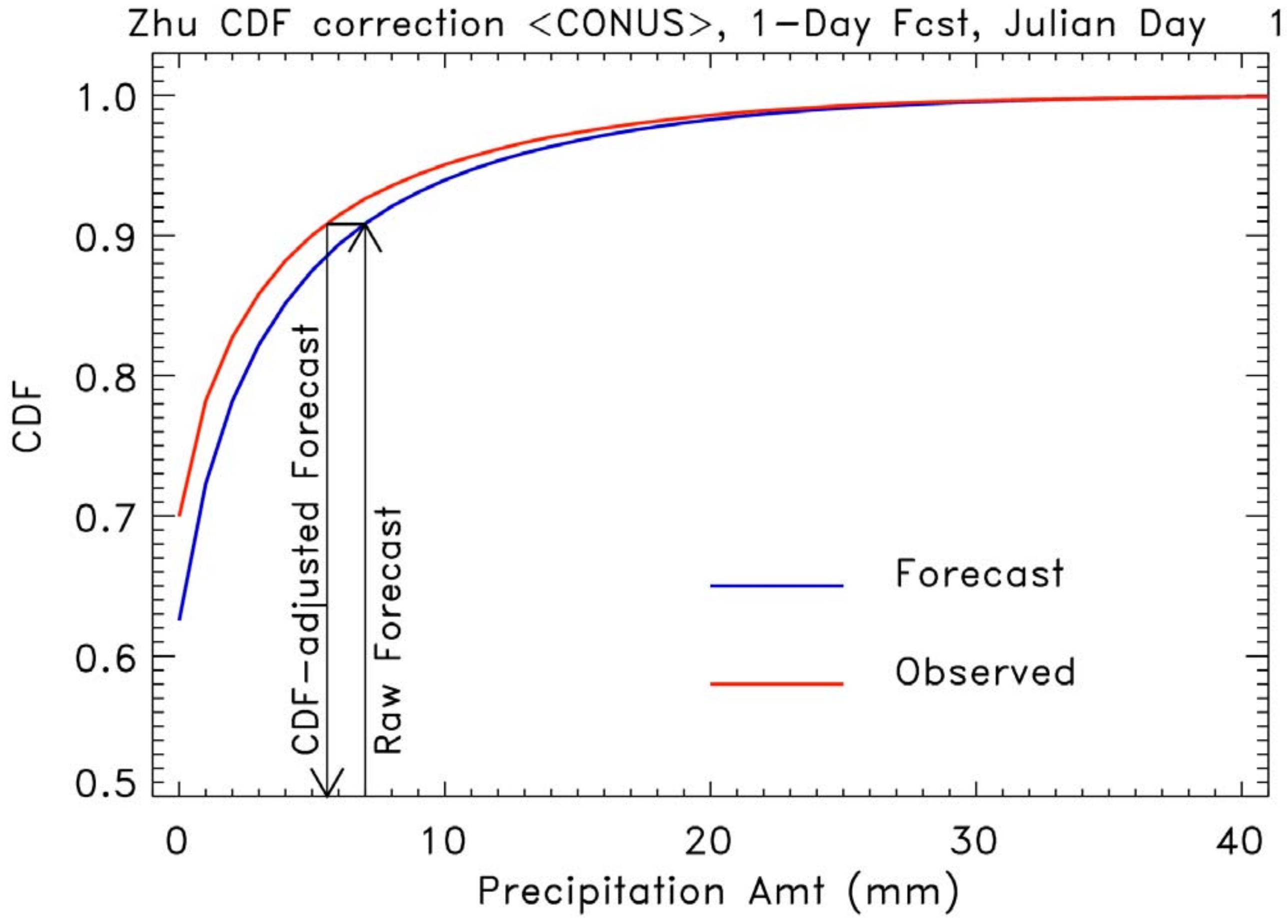
Training sample size, analogs

This shows skill of precipitation forecasts using a two-step analog technique, JFM 1979-2003 data over conterminous US (CONUS). Observations at ~30 km grid spacing (North American Regional Reanalysis).

Notice increased sample size important for calibrating rarer, high-precipitation events.



Sampling issues in other calibration methods: Example “Zhu” NCEP technique



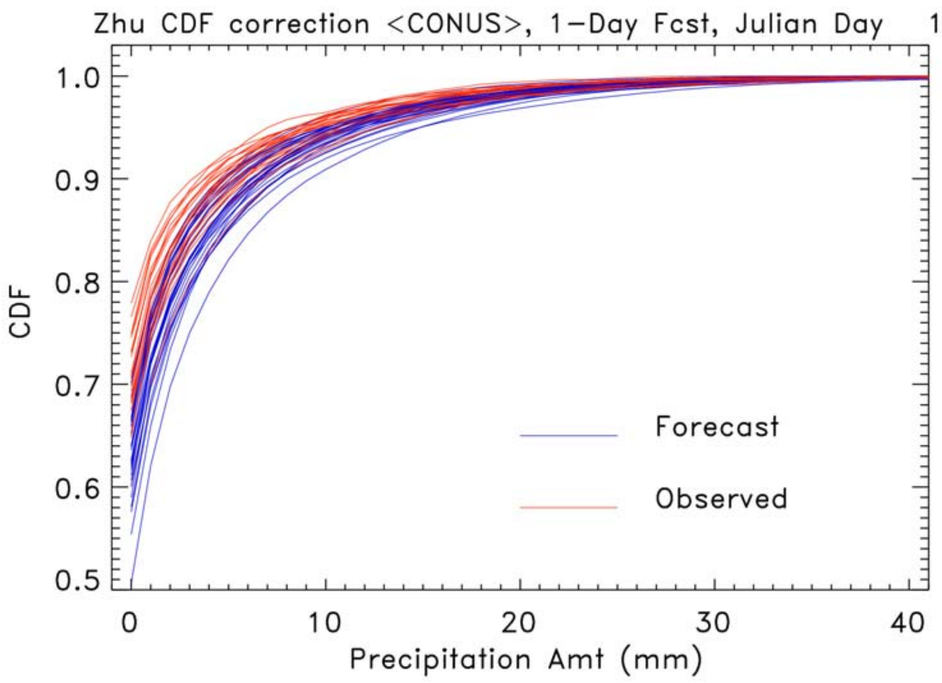
(1) Get CDFs of forecast and observed, averaged over CONUS using, say, last 30 days of data.

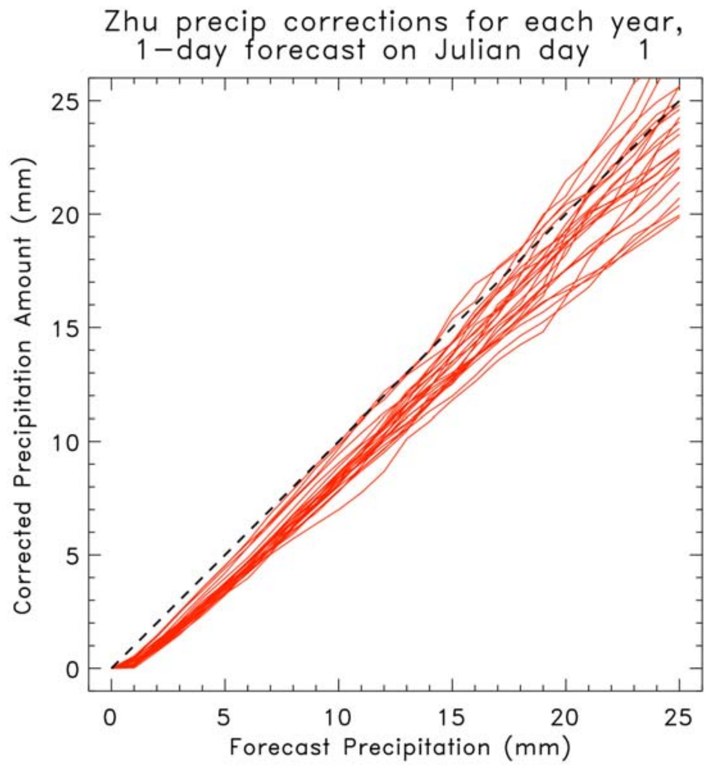
(2) Use difference in CDFs to correct each ensemble member’s forecast. In example shown, raw 7 mm forecast corrected to ~5.6 mm forecast.

NOTE: bias only, not spread correction.

Ref: Zhu and Toth, 2005 AMS Annual Conf.

How much do CONUS-averaged curves vary year by year ?

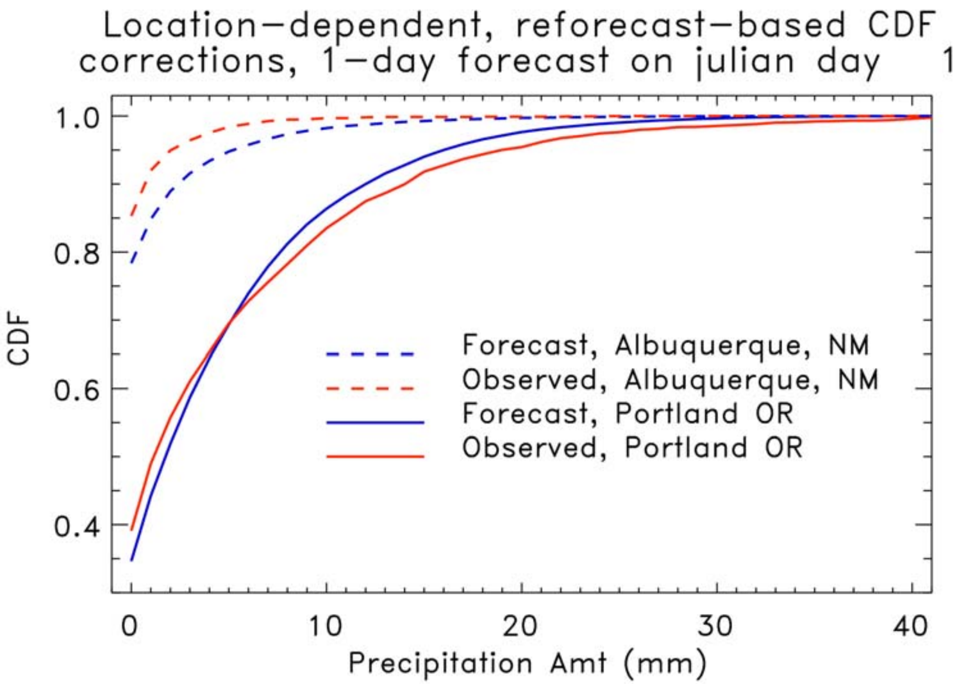




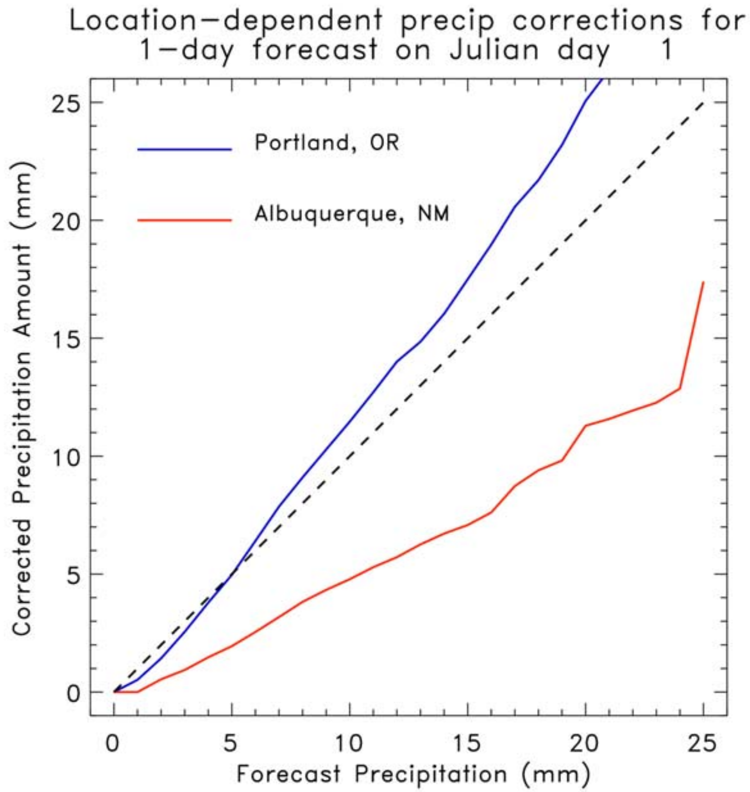
Using reforecast data,
25 different year-to-year
curves are shown above

Each curve shows a year's
corrected forecast amount
as a function of the raw
forecast amount.

What is lost in agglomerating CDF data over many locations in Zhu technique?

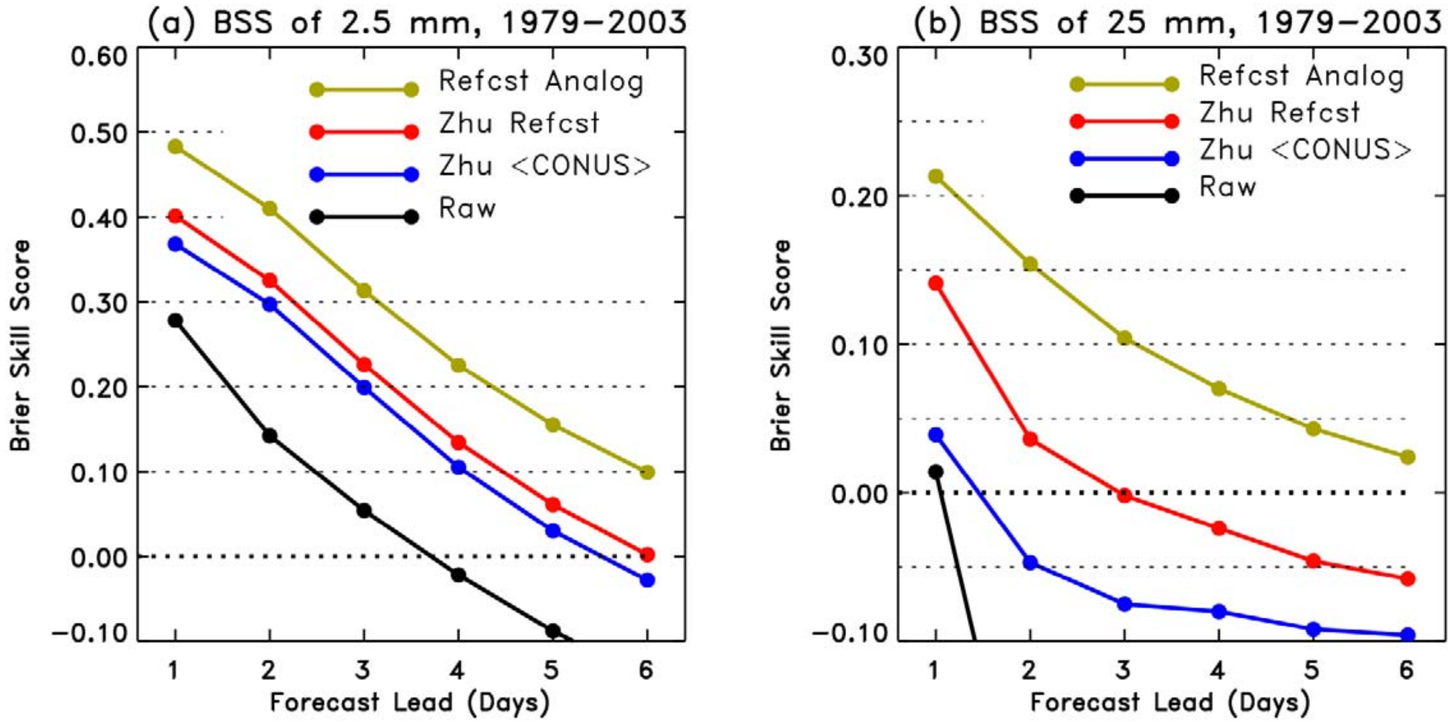


Using the 25 years of reforecasts and a window of +/- 45 days around date of interest, separate CDF estimates were developed for each model grid point. Here are CDFs for two locations on Jan 1.



Different grid points may require dramatically different corrections.

Skill for various precipitation calibration techniques

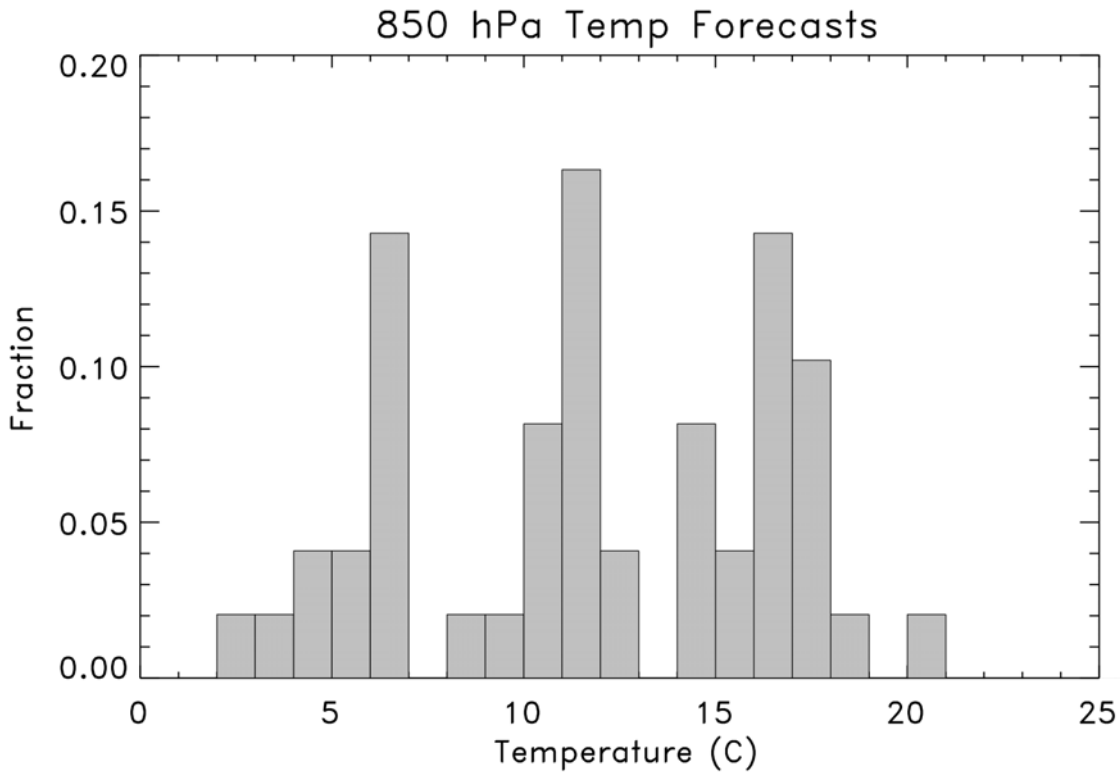


Notes: (1) Here, verification on coarse 2.5 degree grid.
(2) Zhu <CONUS> has benefit at 2.5 mm, correcting drizzle over-forecast.
(3) Location-dependent Zhu technique using reforecasts adds skill, esp. at 25 mm.
(4) Large additional skill by using analog reforecast technique, again largest at high thresholds.
(5) The type of calibration technique really matters.

more

Issues arising in calibration

(2) If ensemble forecasts appeared to be sampled from non-parametric distribution, (e.g., bimodal) should calibration preserve this?



see also Jewson (2004), oai:arXiv.org:physics/0310060

Question: are T850 forecast temperature PDFs normally distributed?

- Test:
  - Generate n=15 random samples from N(0,1)
  - Extract n=15 850 hPa 4-day forecast temps over CONUS.
  - For both random and real data, generate D_n statistic relative to normal distribution fitted to the data, as in “Lilliefors” test.
  - Repeat.

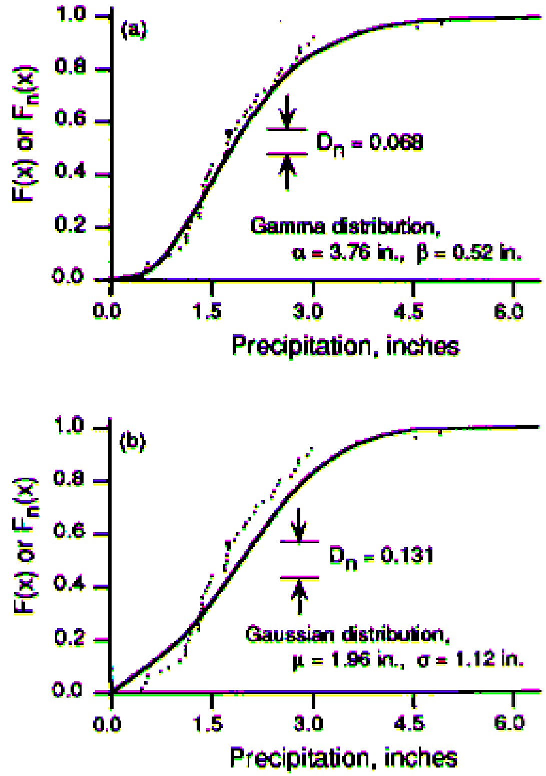
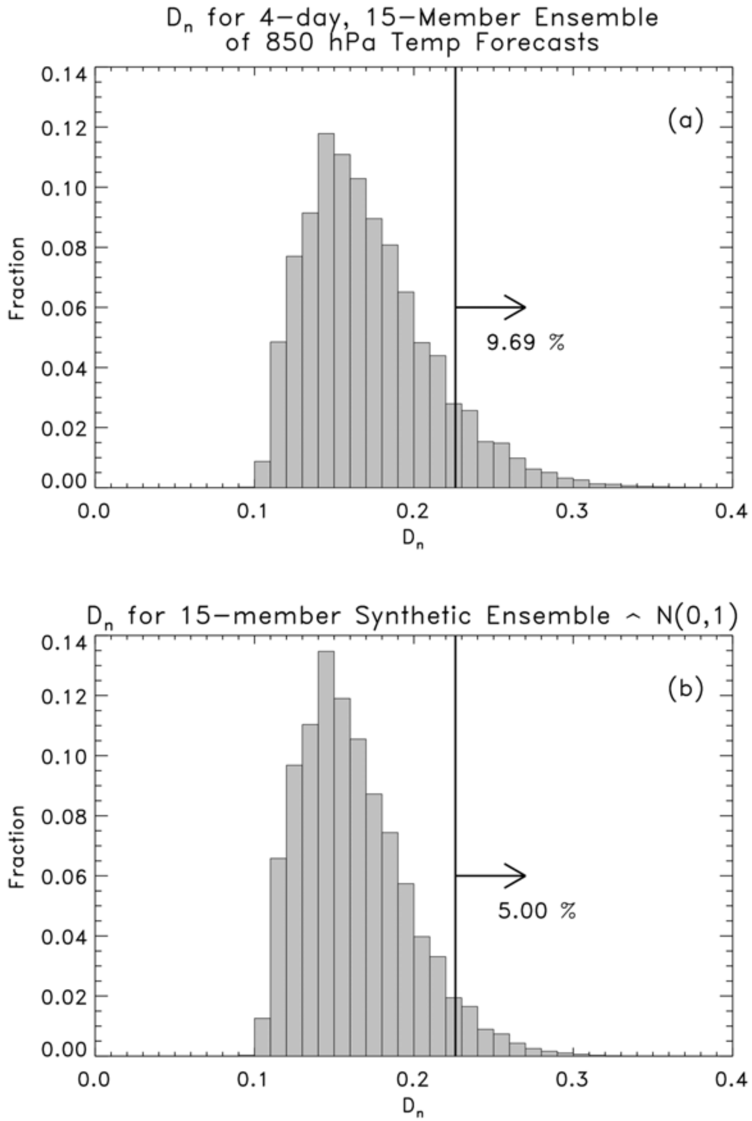


Fig. 5.3 The Kolmogorov-Smirnov D_n statistic as applied to the 1933–1982 Ithaca January precipitation data (dots), fitted to a gamma distribution (a) and a Gaussian distribution (b). Solid curves indicate theoretical cumulative distribution functions, and dots show the corresponding empirical estimates. The maximum difference between the empirical and theoretical CDFs occurs for the highlighted point, and is substantially greater for the Gaussian distribution.

ref: Wilks (1995)

Deviations from normality rare (for T850)



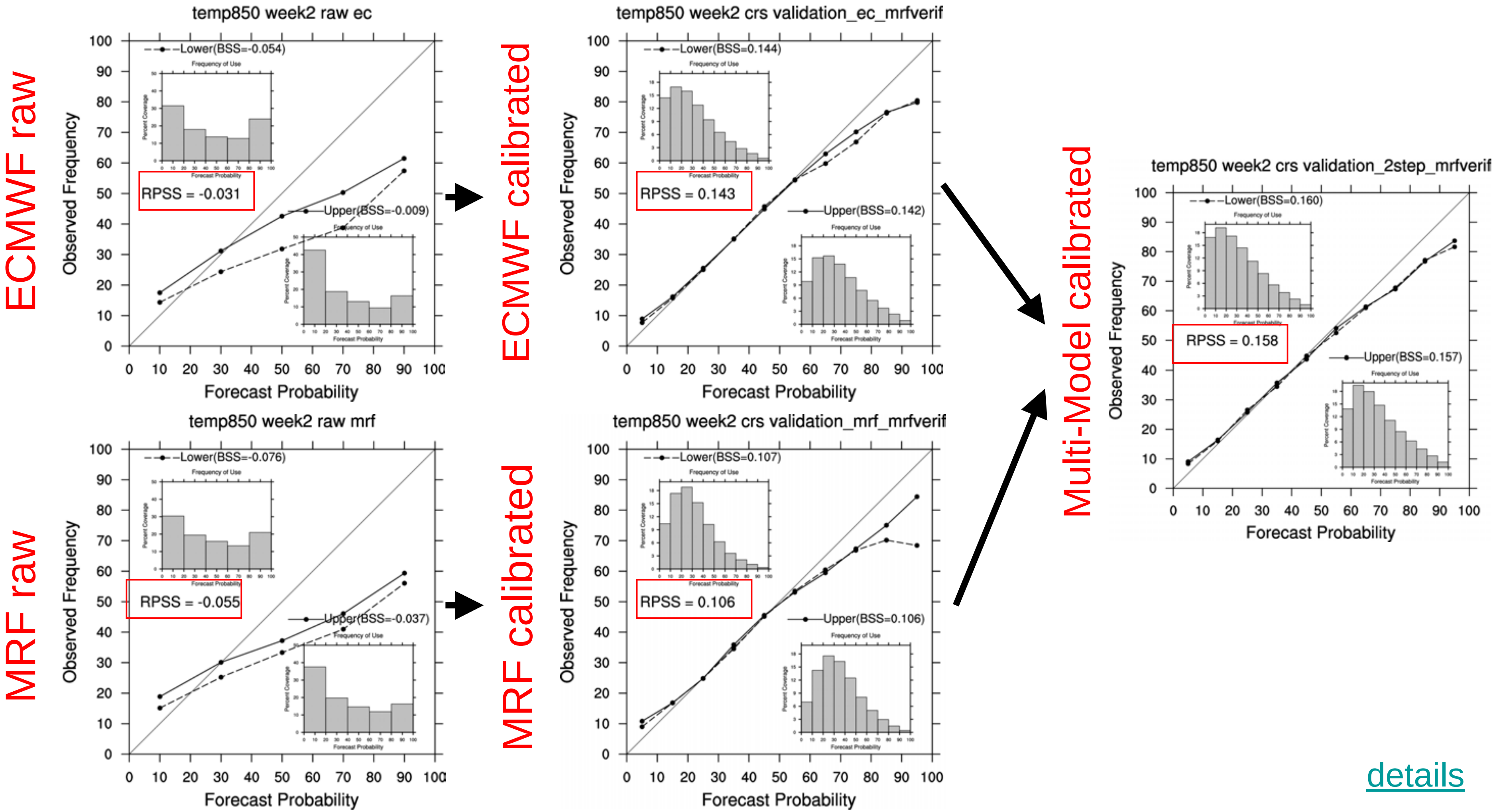
Only 4.69 % of 4-day forecasts have D_n statistic that would justify use of fitting non-normal distribution.

i.e., it's possible you'll more do harm than good by fitting more complicated non-parametric distributions (examples: my old rank histogram techniques, possibly Bayesian Model Averaging).

Lesson: test simple calibration techniques alongside more complex ones.

Issues: (3) Is calibration less necessary when EPS is much improved ?

ECMWF produced 5-member reforecasts once every 2 weeks for 10 years in DJF. Apply logistic regression to ECMWF, CDC, and both for week 2 terciles.



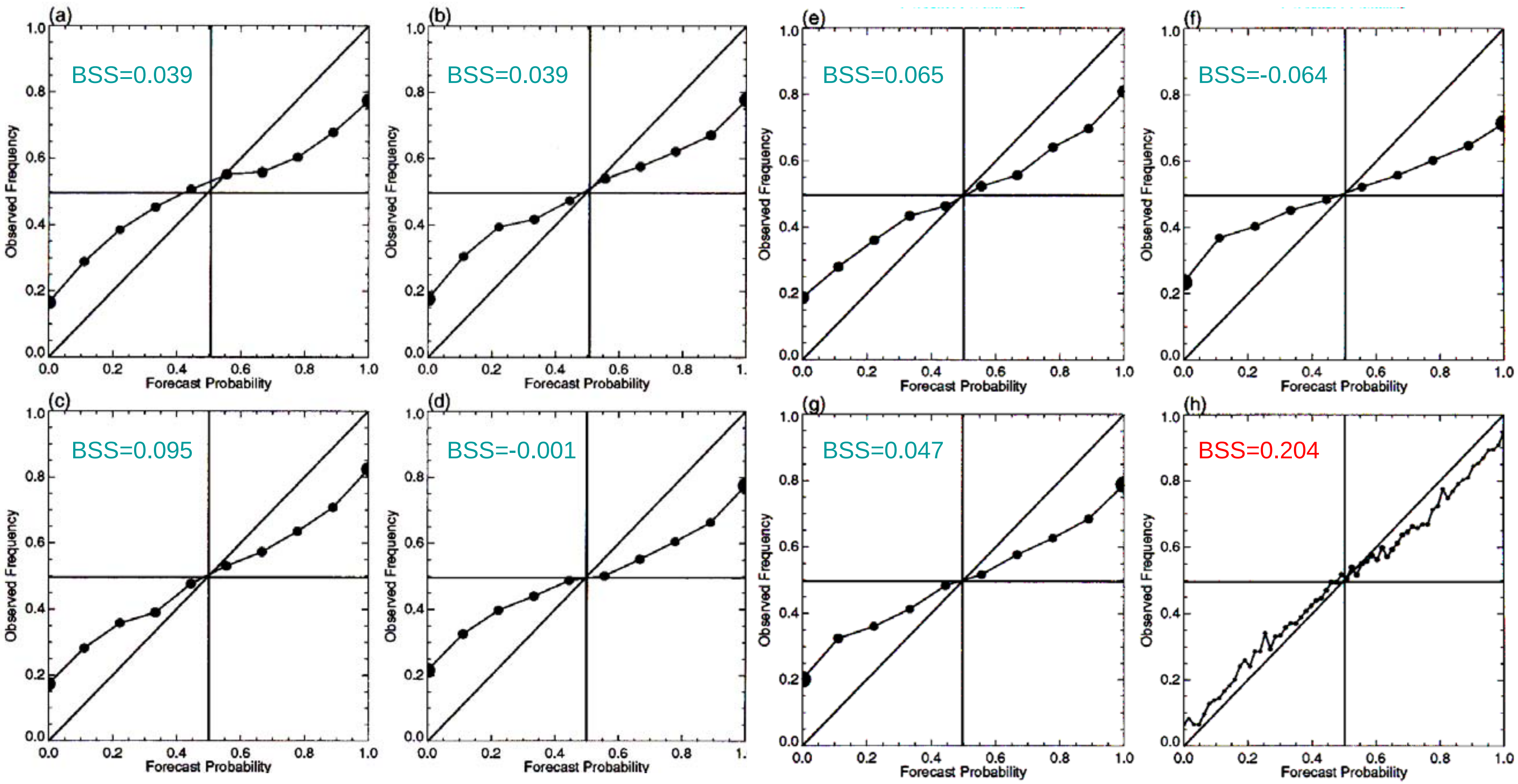
details

Combination: selected multi-model references

- Krishnamurti et al. “Superensemble” : 1999, *Science*, **285**, p. 1548. Multivariate linear regression of multiple models with short training data set improves deterministic forecasts.
- Evans et al. 2000, *MWR*, **128**, p. 3104: Joint UKMO/ECMWF ensembles outperformed either individually. More than bias cancellation.
- Richardson, 2000: *QJRMS*, **127**, p. 1847. Most of benefit in multi-model EFs came from multiple analyses.

Combination of ensembles: the lessons of DEMETER

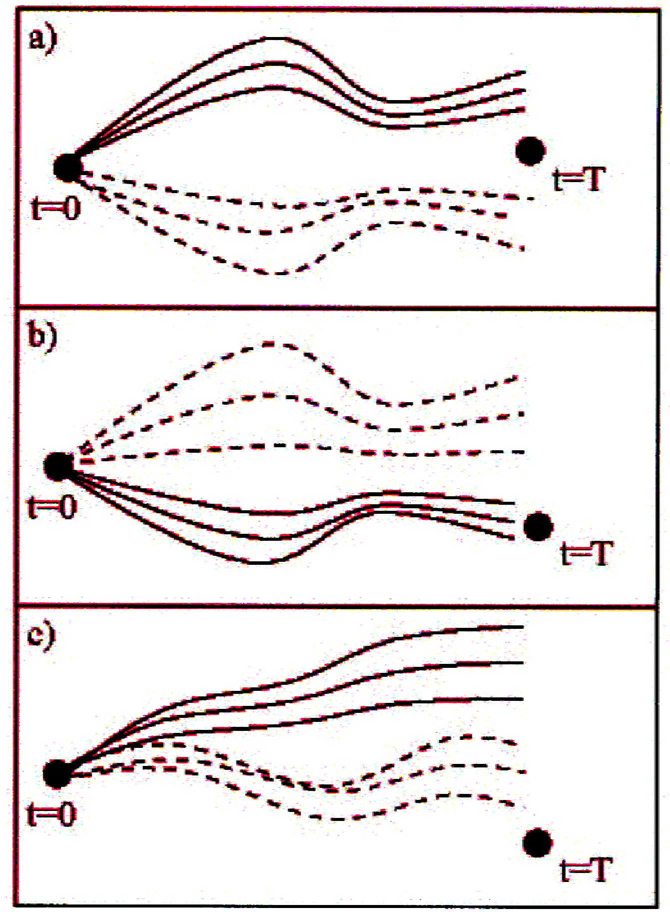
Summer tropical 2-m temp positive anomaly, 1-month lead,
ensembles from ECMWF, CNRM, UKMO, MPI, INGV, LODYC, CERFACS



Ref: Hagedorn et al., *Tellus*, in press, and
www.ecmwf.int/research/demeter

multi-model

Context for multi-models



Observed state within span of multi-model ensemble, not within span of individual ensembles. BIG BENEFIT to multi-model.

One model much more accurate than the other. Might as well rely on the more accurate one.

Both models biased. Multi-model not likely to help much.

Which of these applies for difficult problems like extreme QPF?

ref: ibid.

Potential economic value of DEMETER forecasts

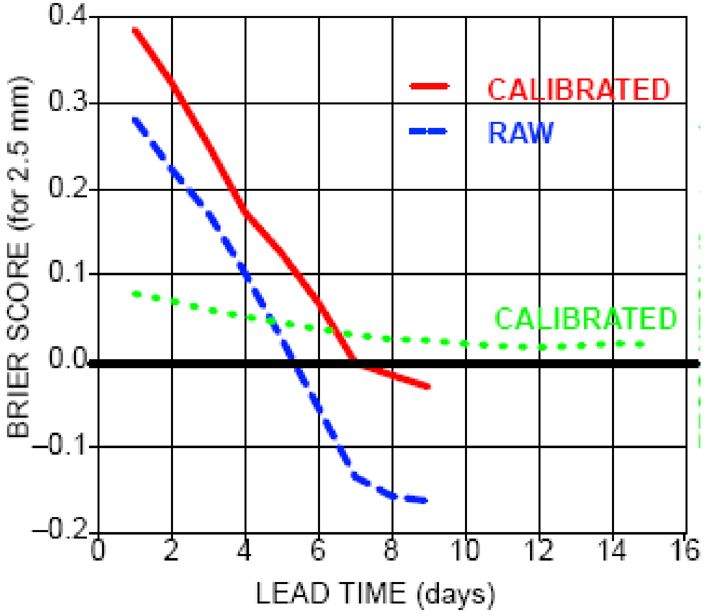
QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Conclusions

- Long reforecast data sets very valuable for calibration. Large skill improvements, especially for rare events.
  - Short training data set → calibration shortcuts →small skill improvement
- Simple, parametric calibration methods should be tried alongside more complicated ones.
- CDC reforecast data set available for your exploration of reforecast techniques. Should be part of TIGGE, too.
- Hope other facilities will explore reforecasts, make theirs part of TIGGE.
- How to reforecast without operational impact? Perhaps do them at reduced resolution, only every few years.
- Encouraging results from preliminary multi-model SREFs and multi-model climate forecasts.

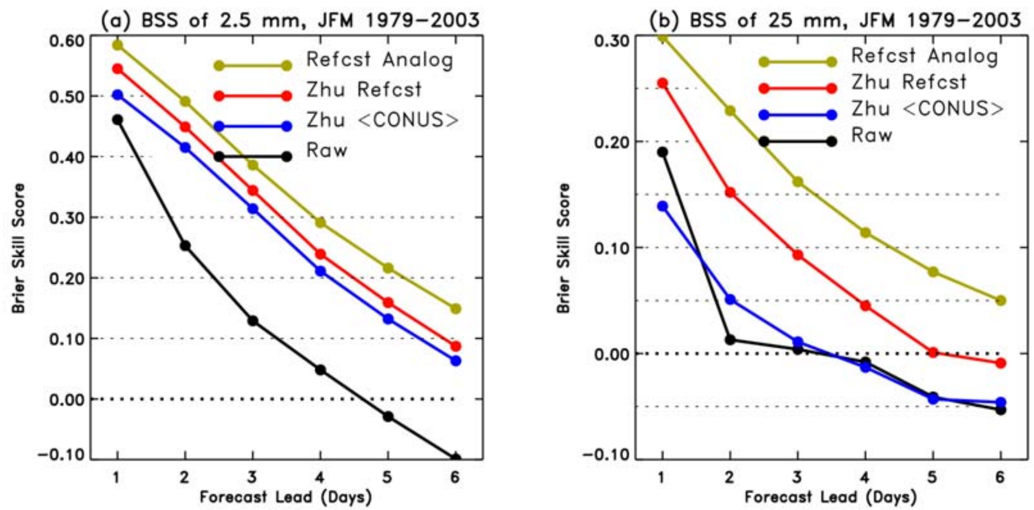
NCEP GFS vs. CDC ensemble

1 Dec 2000 - 28 Feb 2001



(from Zhu's AMS 2005 presentation)

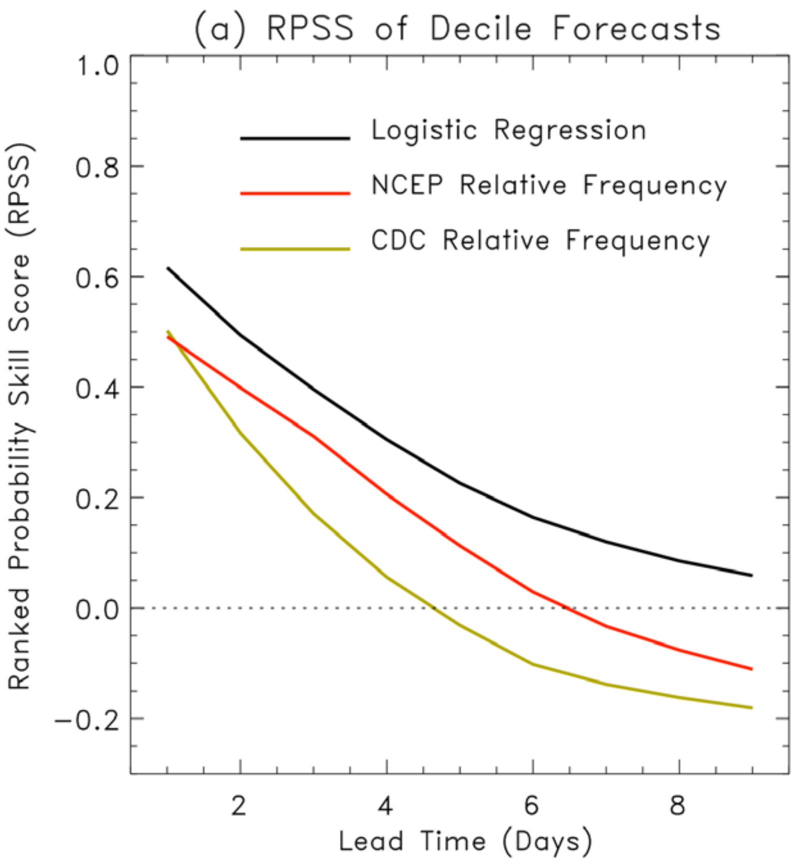
Using CDC reforecast data set, winters 1979-2003

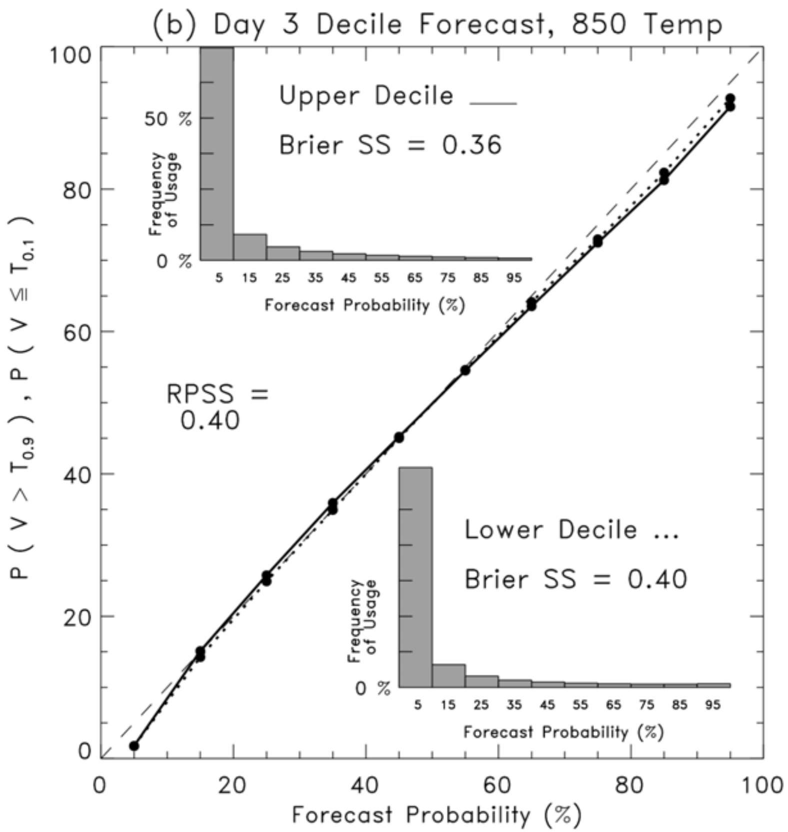


back

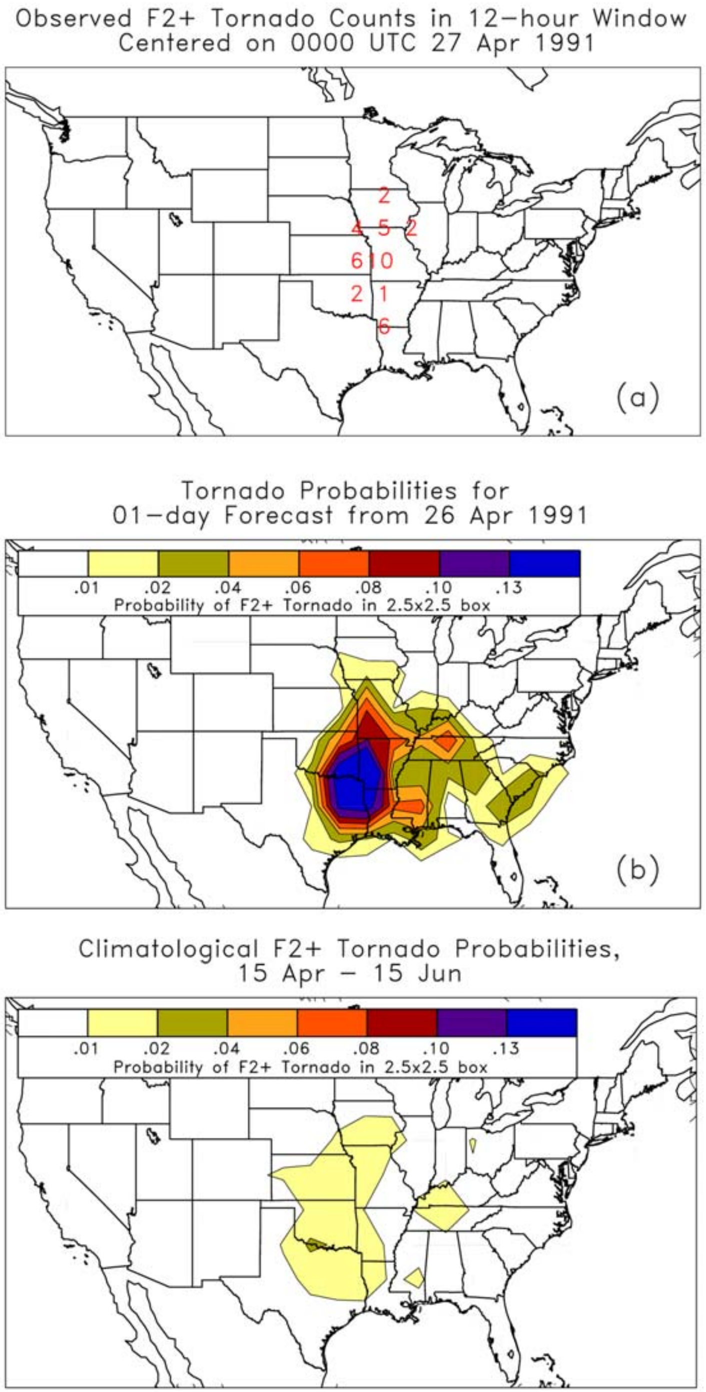
Other examples of calibration using reforecasts

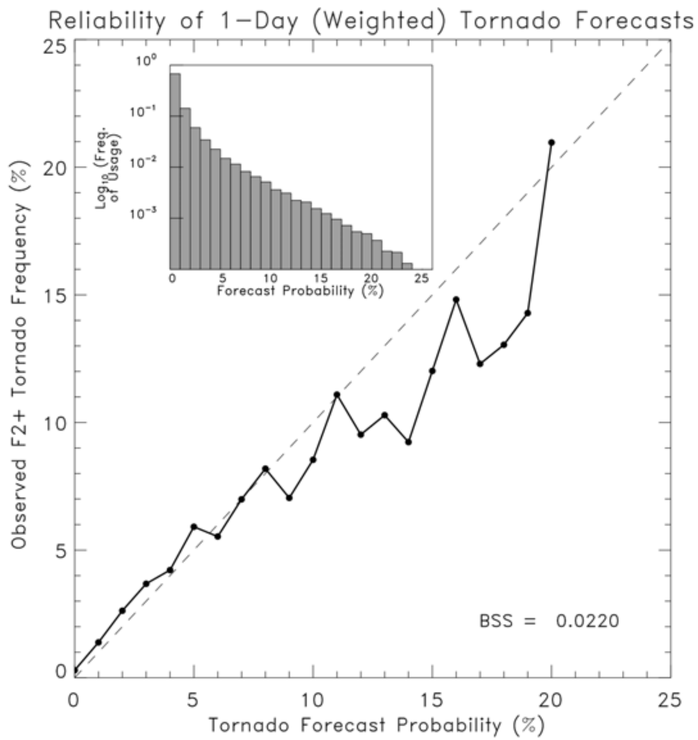
Example: Decile forecasts of 850 hPa temps over US





Tornado Probability Forecasting

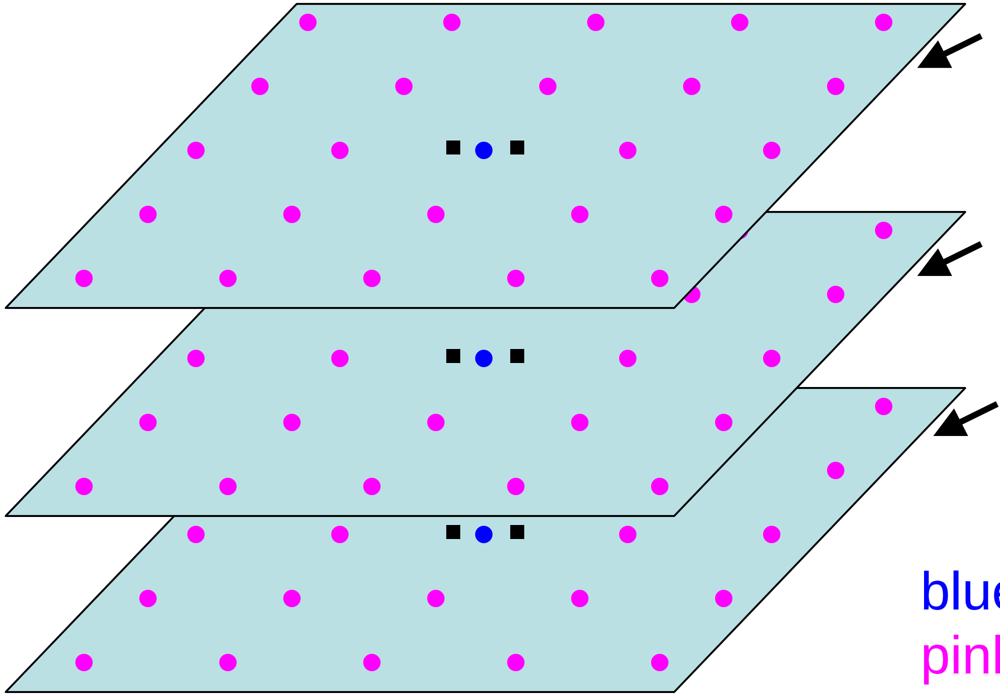




more

Technique for finding tornado forecast analogs

For a given grid point, match today's *ensemble mean* fields with past forecast fields. Find *n* closest analog dates.



Shear / 25 ms^{-1}
(normalized so CAPE doesn't overwhelm shear)

CAPE / 3000 J kg^{-1}

CIN / 500 J kg^{-1}

blue dot: point to find analogs for
pink dots: points to match up current forecasts with old ones.
Note: can vary weights horizontally

Result:

1) Dates of *n* analogs
2) Numerical quantification of how good the pattern match is for each of *n*.

back

Logistic regression with ECMWF and CDC reforecasts

- Forecasts every 2 weeks, DJF for 10 years (85 cases)
- NCEP-NCAR reanalysis for tercile definition.
- CDC, ECMWF separate: run logistic regression on ensemble mean, cross-validated.
- Together:
  - Step 1: Weighted combination of ensemble means
  - Step 2: Logistic regression.
- Details on logistic regression in Hamill et al., *MWR*, **132**, p 1434.

back